（ 続紙 1 ）

京都大学	博士（ 人間・環境学 ）	氏名	渡部崇
論文題目	Regret analysis of constrained irreducible MDPs with reset action （リセット行動が存在する制約付き既約 MDP に対するリグレット解析）		

（論文内容の要旨）

本論文は、離散時間有限状態・有限行動集合での制約付き強化学習の確率的学習アルゴリズムを提案し、そのリグレット解析を行ったものである。強化学習の場合のリグレットは従来の研究で定義されている。さらに、リグレットの累積値が時刻に対し劣線形な函数で高確率で抑えられるアルゴリズムが存在することが既に示されている。しかし制約付き強化学習の場合、まずリグレットの定義が与えられていなかった。本論文ではまずその定義を新たに与えている。加えて学習アルゴリズムを新しく与え、定義に基づいて制約付き強化学習の場合にも制約なしの強化学習の場合と同様の結論が成り立つという新しい解析結果を本論文は与えている。ただし、強化学習の場合にはMarkov Decision Process(以下MDP)の直径というパラメータを函数の引数に含めるなどするのに対し、本論文では異なる条件として、Constrained MDP(制約付き MDP、以下 CMDP)が既約であること、どの状態からも特定の状態に一定以上かつ 1 未満の確率で遷移するリセット行動の存在、与えられた制約を満たす方策が存在することなど、後述の条件を仮定している。

以下では各節における内容を記述する。

第 1 節：導入ではCMDPと MDP の関係、強化学習におけるリグレット解析について大雑把に説明し、論文全体の構成について記述している。

第 2 節：準備ではまず離散時間有限 MDP の定義、方策、定常方策、具体的な最適化の目的函数、それらに纏わる他の概念や函数の定義を与えている。次に強化学習のアルゴリズムの PAC(Probably Approximately Correct)性、時刻ごとのリグレットについて定義を与えた後、CMDPを定義してその場合の対応概念や函数を定義している。またCMDPが既約かつ既知である場合に制約を満たすうちでの最適定常方策を導く線形計画問題を示している。

第 3 節：制約付きUCRLアルゴリズム（以下、CUCRL)では、まず本論文で仮定している 5 つの条件を述べている：1. CMDP が既約であること、2. 報酬函数 R が決定的なこと、3. リセット行動の存在、4. 制約を満たす方策の存在、5. 状態集合 S、行動集合 A の濃度、R、リセット行動とリセット先の状態、制約が既知であること。次に CUCRL を疑似コードで示し、その中で解く必要がある線形計画問題を示している。この問題が厳密に解けることと、実数値が厳密に表され、計算されることを仮定している。CUCRL は制約を満たす定常方策が存在しないと推定される場合には停止し、それ以外の場合には永久に動作する。定理 3.1 と 3.2 は本論文の主定理である。それらの中で制約付き強化学習の場合のリグレット（の時刻 T までの累積値）の定義を与え

ている。本論文の方法では報酬ごとに確率変数として定義される。 Tに対し劣線形な具体的函数を与え、それが高確率でリグレット累積値すべての上界となるというのが定理の主張である。

第 4 節では第 3 節中の線形計画問題と、 CUCRL の動作中の学習におけるエピソードごとの方策の解析を行う。（ 制約付き ）強化学習ではそれまでに得られた経験から(C)MDPのパラメータを確率的に推定することが可能である。そのため、その時点での推定に関する信頼区間を定義し、いくつかの補題と補助的なアルゴリズムを与えて証明を行う。それらのうち補題 4.3 は、直観的な言い方をすると、確率的にリセット動作を行うことで、初期状態による報酬の総和の違いの期待値が抑えられることを示すものである。結果的にこれは、過去研究における MDP の直径に代わる役割を果たすものと考えられる。

第 5 節は主定理の証明を行う。まず、 CUCRL の動作中のエピソードごとに、各時点で推定されているCMDPのパラメータ全体の信頼区間に、真のCMDPがそれまでずっと含まれている確率の下限を与える。含まれていれば第 3 節の線形計画問題は解を持ち、 CUCRL はその時点で停止しない。次にその条件下でリグレットごとに誤差の確率的な上界を評価する。評価はエピソードごとに分解した上でそれらを合計することで行う。最終的に全てのリグレットが上界以下となる確率の評価を行う。

第 6 節：結論は本論文で得られた結果をまとめている。

（続紙 2 ）

（論文審査の結果の要旨）

本論文の主要成果は以下の 2 点である。まず、本論文は制約付き強化学習におけるリグレットの定義を新たに与えた。これは強化学習において既存研究で用いられてきたリグレットの概念を制約付き強化学習の場合に自然に拡張するものとなっている。実際、報酬の数が 1 である場合を考えてみると、定義は強化学習の場合と一致する。また今後の研究において他の学習アルゴリズムについても利用できる汎用的な定義となっている。次に、制約付きUCRLアルゴリズム（以下CUCRL)という確率的学習アルゴリズムを開発し、 CUCRL の場合にリグレットの累積値が時刻 T の劣線形函数で確率的に抑えられることを証明した。制約付き強化学習が問題として学習可能であることを理論的に示す結果の 1 つとなっている。

制約付き強化学習とは、強化学習の拡張の 1 つである。

まず強化学習について説明すると、環境から観測される状態に基づいてエージェントが行動を確率的に決定し、それに対して得られる離散時刻 t ごとの報酬 r_t の重み付き総和（の極限値）の期待値を最大化することを目的とする。期待値を最大化するような（観測された状態からの）行動の決定方法を最適（定常）方策と言い、それらの 1 つ π^* の近似値を求める。環境を有限MDP(Markov Decision Process)とし、報酬が有界であるという条件下では最適となる決定的な定常方策が 1 つ以上必ず存在することがわかっている。ここで有限 MDP とは状態集合と行動集合が共に有限な MDP である。方策が決定的というのは観測された状態に対し一意に行動が定まる方策のことである。定常方策は時刻 t によらず行動の選択確率が一定である方策を指す。

強化学習におけるエージェントは MDP のパラメータとしては状態集合と行動集合の濃度を与えられるだけであり、パラメータあるいは他の行動を決めるためのデータは報酬を含めた実際の観測結果から学習する。そのため、最初から最適方策を取ることはできず、学習にはある種の試行錯誤が必要で、最適方策が分かってから振り返るともっと良い行動があったということになる。学習アルゴリズムの良さを表す方法として、 PAC(Probably Approximately Correct) 性の考え方を強化学習の場合に当て嵌め、高確率で試行錯誤の数の上界となる函数を与える方法と、最初から最適方策を取った場合の期待値との報酬差をリグレットと言って時刻 T までのそれらの累積値の確率的な上界となる函数を与える方法がある。本論文との関連性が強い後者について述べると、強化学習の場合には、 MDP についての一定の条件下で、 T の函数として劣線形、即ち T で割ると $T\to\infty$ の時 0 に収束する函数でリグレットの累積値が高い確率で抑えられる学習アルゴリズムがいくつか知られている。それらの事実はそのアルゴリズムが最適方策を実質うまく学習できることを表している。

制約付き強化学習では、環境は MDP に代わり CMDP(Constrained MDP) で表される。 MDP との違いは主な報酬の他に 1 つ以上の副報酬が存在することであり、副報酬をある一定以上確保しつつ主報酬をできるだけ得ることを目的とする。ただし本

論文での議論は副報酬が 0 個の場合、即ち MDP の場合にもそのまま当て嵌まると考えてよい。本論文では状態数有限等の上の MDP の条件に、CMDPの既約性を加えた条件下で、i 番目の副報酬の無限平均の期待値が c_i 以上という条件を満たしつつ主報酬の無限平均を最大化することを目的とする。制約付きだと制約条件を満たす方策が存在しない場合、あるいは存在しても決定的なものは存在しない場合があり、強化学習の場合とはある程度以上異なる問題になっていることがわかる。

本論文は制約なしの強化学習のリグレット解析を制約付きの場合に拡張する方法を与えている。実際に具体的なアルゴリズムを与えた上、それについて、本論文で新たに定義されたリグレットの累積値が確率的に劣線形函数で抑えられることを示したものであり、理論面で制約付き強化学習の新たな地平を開くものと言える。

以上により本論文は博士（人間・環境学）の学位論文として価値あるものと認める。また、令和 2 年 1 月14日、論文内容とそれに関連した事項について試問を行った結果、合格と認めた。

要旨公表可能日：　令和　　年　　月　　日以降